Publishing & Editing Datasets

A DATAshare Guidance Document

Iowa Department of Management
10/14/2013

Introduction

Now that you have selected and organized your data, the next step is to ensure the public will be able to understand your data. To achieve this, you need to give it context. When people know who, what, when, where, why, and how – they are better able to interpret your data, and see its value.

This guide is intended to provide information on:

- How to provide necessary context for your dataset,
- How to place controls on your data to prevent misrepresentation, and
- How to publish your data in DATAshare.

Contents

Where do I go to create a new dataset?

When you sign in you will have access to "My Workbench" on the right side of DATAshare's home page, as shown in figure 1. One of the choices available to you, depending on your privileges, is "Create Dataset." When you click this link you will be taken to a web form that begins the process of publishing a dataset, which includes drafting information about the dataset, and configuring the dataset's columns.

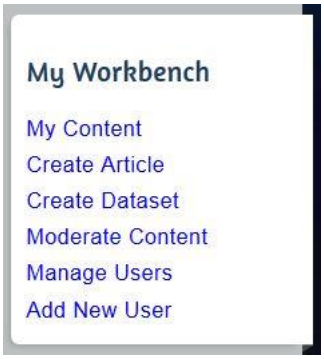

Figure 1. My Workbench is available in the right hand side of the home page. Links available vary depending on user's role.

If "Create Dataset" is not available to you when you sign in, it is because you have not been assigned to a creator or moderator role. Contact the Department of Management to be assigned a role.

How should I title my dataset?

While we do not have required naming conventions for datasets in DATAshare, it is important to provide a name that allows users to quickly determine the type of information provided in your dataset. Some basic elements should be considered when coming up with a title for your dataset:

- The main numeric data available within your dataset should provide the foundation for your title (e.g. Vendor Payments; Assessed Property Values; Local Option Sales Tax Rates & Payments)
- Known timeframes your dataset is limited to should also be used if applicable (e.g. FY 2013 Vendor Payments; 2012 Assessed Property Values)
- Groupings used to summarize underlying data where record level detail is either not available or not provided due to sensitive or confidential data (e.g. 2012 Assessed Property Values by Tax District; FY 2014 Monthly Medicaid Payments by Vendor)

What should I include in the dataset's summary?

The dataset's summary is where you provide context for your dataset. The summary is intended to describe the type of information within a dataset, its use limitations, and completeness. Your narrative should answer the following questions:

- What does my dataset help explain or describe?
- Why is it important to my agency?
- Why was the data collected?
- How was the data collected?

- Are there uses that the data should be restricted to or discouraged from (e.g. limitations)?
- Is there some subset of the data that is missing (e.g. completeness)?

You will want to ensure your narrative is easily understood – appropriate to the public's reading skills, and knowledge. It should also be clear and direct, free of unnecessary jargon, acronyms and abbreviations. Oftentimes acronyms and abbreviations have multiple meanings in different areas of government, industry, or even walks of life. As such, unintended meanings for abbreviations and acronyms use can cause confusion and uncertainty in what your data conveys. Here are some tips for writing the summary:

- Avoid large words
- Limit or avoid use of jargon, acronyms, and abbreviations
- Use headings, bullets, and numbered lists to guide the reader
- Use present tense
- Use active voice
- Remember the general public should be able to understand what you have written

What should I consider when selecting topic tags?

Topic tags facilitate searches in DATAshare's catalog, so you should take some time to consider the most appropriate ones to use. Below are items you will want to consider when identifying topic tags:

- Identify topic tags already used on datasets providing related or similar information
- Select topics tags that describe the main numeric data you intend the public to summarize
- Ensure topic tags highlight ways in which the data can be categorized
- If your dataset is associated with a specific timeframe (e.g. FY 2013, CY 2013, etc.), use it has a topic tag
- If your dataset is associated with a specific geographic region (e.g. Polk County, Des Moines, etc.), use it as a topic tag

Before you begin creating your own topic tags, consider selecting from existing topics which are provided when you start typing at least three characters. If you draft your own, please consider the following:

- Don't use more than two words in a topic tag
- Make sure you spell your topic correctly
- Consider using upper case characters at the beginning of each word

How do I relate agencies to my data?

You will need to list your agency in the "Provided by" field, see figure 2. This field limits your choices to the agencies you have authorization to represent. The data you publish should be data your agency is responsible for collecting, maintaining and updating.

Figure 2. Relating Agencies to the Data

You are also able to specify one or more agencies that are represented in the data. The organization taxonomy is hierarchical based, so you can be as specific as needed. Agencies should only be tagged if the data is about them in some way – such as related to the agency's administration (e.g. budgets, expenditures, etc.) or the agency's performance. If there are a lot of state agencies represented in a specific dataset, you can select "Comprehensive" rather than selecting them individually. You may select more than one by selecting the first, then holding the CTRL button on the keyboard and left mouse clicking on additional organizations.

How do I specify the timeframe associated with my data?

Your data is likely linked to time in some way - it might be a time series, or a snapshot from a specific period. The period of coverage fields on the web form provide a place to inform the public of when your data was collected. It is important to provide this so old data is not presumed to be current data. Unfortunately, it's a common mistake to take old data and pass it off as new because it's what's available. The period of coverage can and should be revised if your data is updated to include more recent information.

How can I limit access to my data?

While we encourage you to make your data fully accessible, you can remove the public's ability to download the entire dataset as a CSV file and disable their ability to create interactive visualizations.

Once you have uploaded your CSV file containing data associated with the dataset. You may check or uncheck the "Include file in display" check box, see figure 3. When this is checked, users have the ability to download a CSV file containing the entire dataset. When it is not, the

data download option is not available on the public display of your dataset.

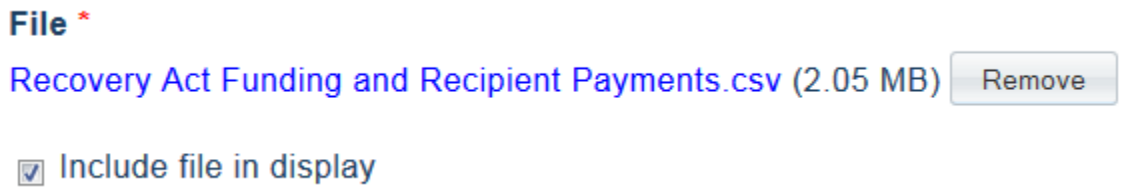

Figure 3. Uncheck "Include file in display" to prevent public from downloading entire dataset

Additionally, on the web form, you have the ability to "Disable Visualization Creation," see figure 4. This allows agencies to prevent the public from being able to create interactive visualizations for your dataset. You may want to prevent public users from creating visualizations if you have a complex dataset that could be easily misrepresented.

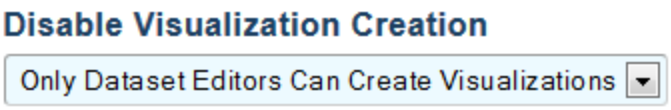

Figure 4. Disable Visualization Creation prevents the public from creating tables, charts and maps.

What should I enter for the column title and description?

The title of each column (field with in the CSV file) can be provided by file where the first record contains headings – which is the preferred method. However, the title can be changed when you are configuring your columns. The title serves as the key label for the column, and should be very brief (i.e. two to three words at most). The title given should be closely related to the data contained in the column.

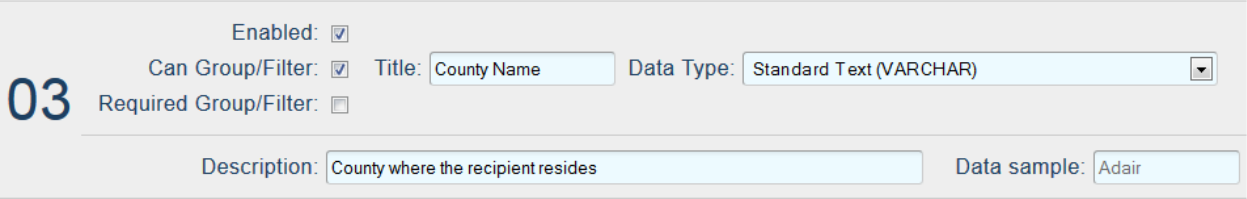

Figure 5. Column title and description fields are available on the column configuration page.

The column description provides a definition for the column so that the public understands what data the column includes. The description should be a brief sentence. Although you may feel the column title is intuitive, the description offers further explanation. For instance, if your column title is "County", you should use the description to state how the "County" data is related to the rest of your data such as, "County where program recipients live." These relationships may not be intuitive to the public who are not as familiar with your data.

What controls are available for my data?

Providing controls over how your data are summarized helps ensure your data are not misrepresented when charts, maps or tables are produced. For example, you are able to turn off row counts for datasets containing aggregated records. You can also turn off calculations available for numeric columns – such as preventing users from summing current age of residents. You are also able to specify which columns can be grouped and filtered when users are creating charts, tables or maps – and if grouping and filtering are required. You may also have concerns about providing granular (detailed) data due to potential identification of individual students, taxpayers, program recipients, etc.

When you have completed and saved the web form containing the summary and other metadata, you are taken to a step where you can configure your dataset's columns (e.g. fields in your CSV file), and implement a number of controls, see Figure 6. The following highlights the available controls in DATAshare:

- Allow Count of Records – when checked, "Count of Records" is available under calculation options when creating visualizations. You will likely want to uncheck this where records within your dataset have been aggregated. If count is allowed, you will need to provide a title and description that will be used in visualizations
- Suppress data smaller than – when a numeric threshold is entered, counts and calculated results of numeric columns (where data suppression threshold is applied) less than the number entered will not be displayed in visualizations, or the download of the

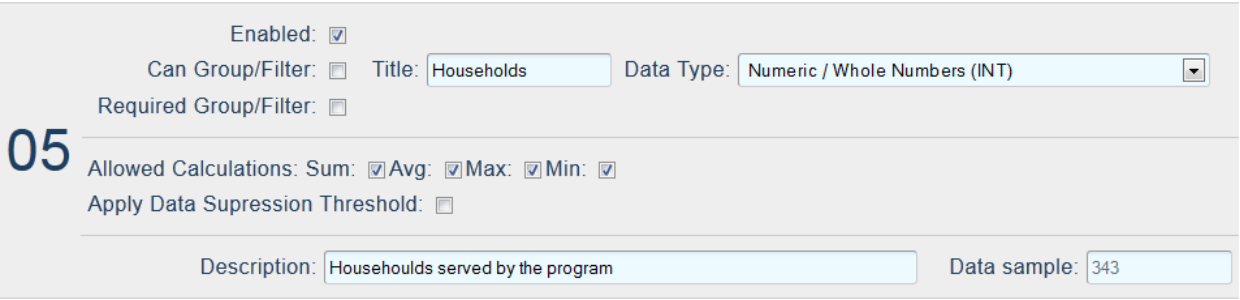

Figure 6. Available configurations for numeric columns (i.e. fields) in CSV files.

visualization's data.

- Enabled – when checked, the column is available for creating visualizations
- Can Group/Filter – when checked, the column is provided as an option for grouping and filtering the dataset when creating visualizations
- Required Group/Filter – when checked, the column has to be selected for filtering or grouping to help ensure the data is not misrepresented

- Allowed Calculations (visible under numeric columns) – when checked, the functions sum, min, max, and average are made available under calculation options when creating visualizations for the column.
- Apply Data Suppression Threshold – when checked, calculated results for the column falling below the threshold specified in “Suppress data smaller than” are not displayed in visualizations or their data available for download.

What data types should I choose?

There are a number of data types, but most columns will either be text, numeric (multiple formatting options), date, or date/time. The data type dictates whether the column supports grouping/filtering and calculations, and how the data will be formatted in visualization displays. There are a number of special data types that represent address information, or codes to geographic regions and for financial and accounting purposes. Many are tied to geographic regions that support mapping (e.g. counties, cities, school districts, townships, zip codes), or are in place to support future join capabilities.

See Appendix A for a complete list of data types.

What are some common problems with importing data?

When publishing your data in DATAshare, it is imported into a database. This allows the system to index your data and make it available for the creation of interactive visualizations. Unfortunately, no importing process works perfectly 100% of the time. There are a couple of the common problems you may encounter:

- Data in column do not match the data type selected (e.g. text in columns designated as a date or numeric data type)
- Text within dataset contains garbage characters (e.g. â€“, â€™, â€œ, â€, ^Å, etc.). This can happen where data are collected through web forms that allow users to copy and paste text from Microsoft Word or other sources and stores it in a database based on a different encoding scheme. Single and double "smart quotes", en and em dash, ligature characters (such as the ampersand (&)) and ellipsis (…) are some characters that can result in garbage characters.

To ensure your dataset is complete, you will need to correct problems encountered during the import process – such as deleting text in numeric columns (or changing the data type for the column), and removing/replacing garbage characters in your CSV file. The import log will help you isolate which records were not imported. Once corrections are made in the CSV file you

will either need to re-import the entire file, or append a file that contains only those records that were not imported originally.

How do I read the import log?

The import log helps you isolate problems occurring during the import process. Once the import process is complete you will be taken to a page that gives you an option to download the import log, see figure 7. You should look at this log when you have imported, re-imported or appended data in DATAshare.

Figure 7. Once you have imported your dataset, download the import log to verify all your data was successfully imported

The import log is a text file that can be opened in Notepad.

How do you read the row and column references?

When you open the text file containing messages about the dataset, you will see references to rows and columns. This information can help you identify what cells within your dataset requires corrections. However, if you open you CSV file in Microsoft Excel, you will notice that the row and column references in the log do not directly correspond to row and column identifiers in Microsoft Excel. The references in the log relate to how the data is indexed. The first row containing data is zero. Rows in the data that contain your column headings is not referenced, see tables 1 and 2 below.

Table 1. Row References in CSV files with Headings

Row Reference in Import Log	Row in Excel
0	2
1	3
2	4
3	5
And so on…	And so on…

Table 2. Row References in CSV files without Headings

Row Reference in Import Log	Row in Excel
0	1
1	2
2	3
3	4
And so on…	And so on…

Column references are done in a similar fashion, see table 3 below.

Table 3. Column References

Column Reference in Import Log	Column in Excel
0	A
1	B
2	C
3	D
And so on…	And so on…

Figure 8 shows how the log can be used to isolate issues with the import. In this example, a row was not imported due to "garbage characters" referenced in the previous section.

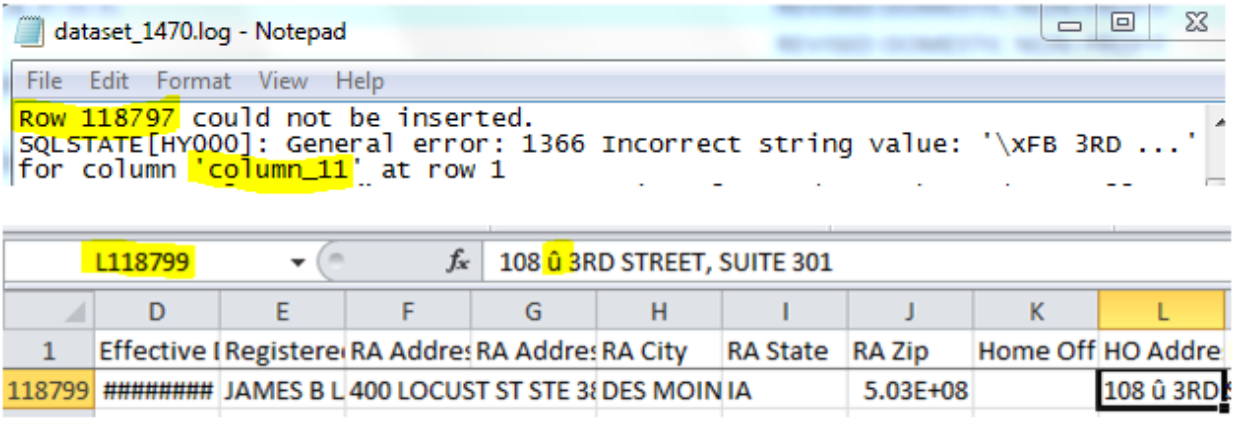

Figure 8. Import Log - Excel Reference Comparison

How do I find information on rows not imported?

If you have a very large dataset, finding information from amongst all of the messages about changing fields with no values to null in numeric columns can be challenging. In Notepad, under the Edit menu option, select "Find." This will open a dialogue box, where you can enter

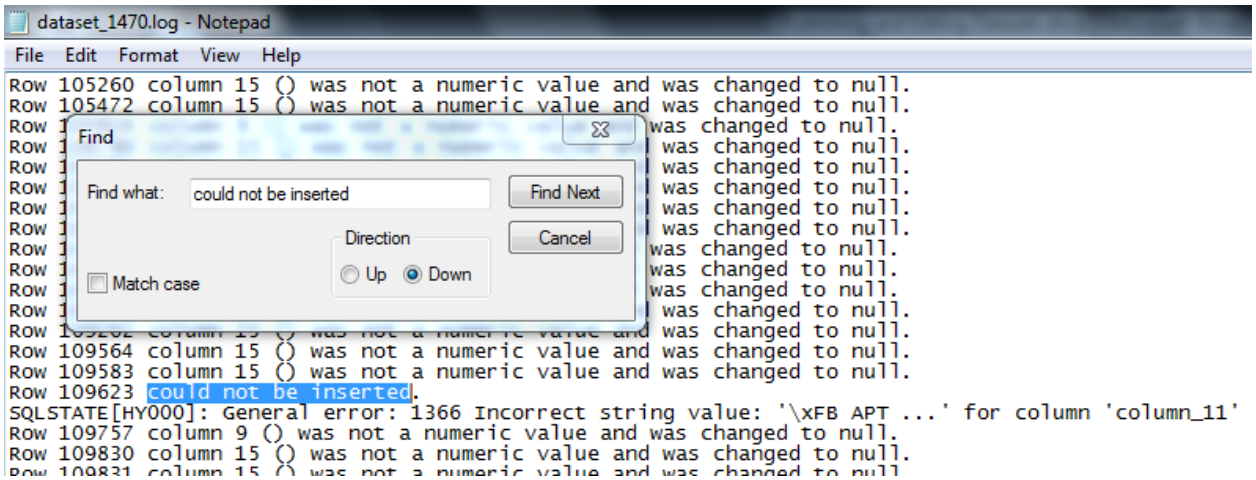

Figure 9. Finding rows not inserted. Go to edit menu option, and select find.

the following text "could not be inserted" then click find next, see Figure 9.

How do I edit my dataset?

Once you have created a dataset, you can easily edit it by accessing your "My Content" page. Once logged in, click "My Content" link available under the "My Workbench" section on the home page of DATAshare, see figure 1. After accessing your "My Content" page, you can filter your content by type, publishing type, and/or terms to quickly find what you are interested in

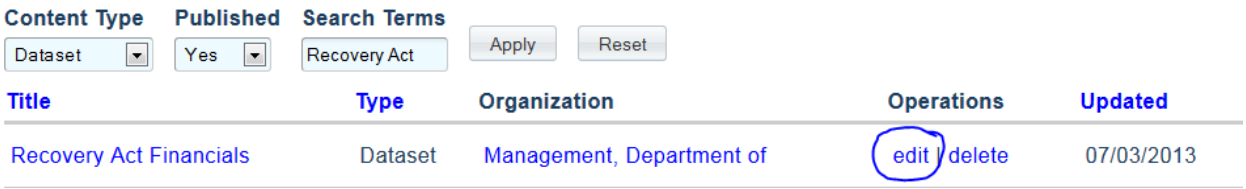

Figure 10. After filtering (if necessary) click edit operation.

editing. Click the apply button to run the filter or reset to clear it. Once, you have found the dataset you are interested in editing, click the "edit" link in the operations column of the table, see Figure 10. This will open the web form that you completed when first creating the dataset.

What do I do to update the data?

After accessing the web form, you will need to attach the file containing the update to your data. You do that by first clicking the "Remove" button available under the "File" section of the

web form, see Figure 3. Once you have done that, browse, open and upload the CSV file containing your update.

Once you have made any changes necessary on the web form, click the "Save" button at the bottom of the form. When you do this you will be presented with some options:

- Reconfigure – Reconfiguring allows you to change your column definitions, titles, or controls you have placed on the dataset. If you select this option, the data you have uploaded will replace the data that currently exists with the data contained in the attached CSV file. Reconfiguring your data could potentially impact existing visualizations, so take care when proceeding with this step.
- Re-import – Re-importing your data will replace your entire dataset with the data contained in the attached CSV file. To re-import your data, it is important that your CSV file is structured exactly like your original CSV file. If changes have been made, you should reconfigure your dataset to address these changes.
- Append Data – Appending your data will append the data contained in the attached CSV file to the data already contained in the dataset. As with re-importing your data, when you append data you need to ensure your CSV file is structured exactly like your original CSV file (or is based on last reconfiguration made).
- Just Publish It/Just View It – This will not update your data but will just reflect changes made to narrative, tags, etc. shown on the dataset display.

How do I publish my agency's datasets?

Those assigned a moderator role in DATAshare will have a "Moderate Content" link on the "My Workbench" section of the home page, see Figure 1. By clicking this link you will be taken to the Moderate content page, which lists all of the content owned by the agency you are assigned to and any suborganizations. You can filter the list a number of different ways, see

Figure 11. Moderate content page allows Moderators to filter, edit, delete, and publish/unpublish content.

Figure 11. Publishing a dataset is as simple as clicking the "publish" link under the operations column of the table. You also have the ability to unpublish content that is currently published. Moderators also have the ability to edit datasets created by others within their agency, and delete them if necessary.

Conclusion

Congratulations for taking the necessary steps to publish your data. Time spent on these steps will help ensure the public understands your data, and is not able to create summaries that misrepresent your data. Before you make your data public, confirm that you have:

- ✓ Created a title that facilitates understanding what your data covers
- ✓ Provided information on the importance of your dataset, an explanation of why the data is collected, and your data's limitations and completeness
- ✓ Highlighted agencies, topics, geography and timeframes that will facilitate the public finding your data.
- ✓ Set necessary controls to help minimize the possibility that the data will be misrepresented.
- ✓ Corrected import problems to ensure your data is complete.

Appendix A. Data Types

Data Type	Definition
Standard Text (VARCHAR)	Text with 255 characters or less
Numeric/Whole Numbers (INT)	Positive or negative numbers without a fractional or decimal component. For negative numbers, the minus sign at the beginning or end of the number is accepted.
Numbers with Decimals (DECIMAL.00)	Numbers with a decimal component displayed to the hundredths. For negative numbers, the minus sign at the beginning or end of the number is accepted.
Numbers with up to 6 Decimals (DECIMAL.000000)	Numbers with a decimal component displayed out to six decimal places. For negative numbers, the minus sign at the beginning or end of the number is accepted.
Dollar Amounts ($DECIMAL.00)	Dollars displayed to the penny. For negative numbers, the minus sign at the beginning or end of the number is accepted.
Whole Dollar Amounts ($000)	Dollars displayed without cents. For negative numbers, the minus sign at the beginning or end of the number is accepted.
Decimal Percentages (DECIMAL.00%)	Percentages expressed as decimals (e.g. 0.12). For negative numbers, the minus sign at the beginning or end of the number is accepted.
Whole Percentages (VALUE%)	Percentages expressed as integer (e.g. 12). For negative numbers, the minus sign at the beginning or end of the number is accepted.
Text with more than 255 Characters (TEXT)	Text with more than 255 characters
Date Values (DATE)	Date in one of the following formats: 12/31/2013 2013-12-31 20131231 Dec 31, 2013 December 31, 2013
Date with Time Values (DATETIME)	Date with a time stamp in one of the following formats: 12/31/2013 14:28:59 12/31/2013 02:28:59 PM 2013-12-31 14:28:59 2013-12-31 02:28:59 PM 20131231 14:28:59 20131231 02:28:59 PM Dec 31, 2013 14:28:59 Dec 31, 2013 02:28:59 PM December 31, 2013 14:28:59 December 31, 2013 02:28:59 PM
Street Address	Physical Street Address
City Name	Name of city, such as those found in a postal address (e.g. Des Moines, not City of Des Moines or Des Moines city).
State Name	Two character state abbreviations used by the United States Postal Service, or the name of the state spelled out (e.g. IA or Iowa)
5 Digit Postal Code	Five digits assigned by the United States Postal Service.

Data Type	Definition
9 Digit Postal Code	Nine digit zip code assigned by the United States Postal Service, also known as zip + four. Accepted Formats: 50319-0001 503190001
Iowa County Number	A two digit numeric identifier for Iowa Counties assigned to counties in alphabetical order starting with Adair County (01) and ending with Wright County (99).
FIPS County Code	County (FIPS Code) is a five-digit Federal Information Processing Series (FIPS) code issued by the National Institute of Standards and Technology (NIST) to ensure uniform identification of counties in the United States through all federal government agencies. The first two digits are the FIPS state code and the last three are the county code within the state or possession.
Iowa DOT City Number	Code given to cities by the Iowa Department of Transportation.
City FIPS Number	City (FIPS Code) is a seven-digit Federal information processing standards (FIPS) codes issued by the National Institute of Standards and Technology (NIST) to ensure uniform identification of cities and other populated places through all federal government agencies. The first two digits are the FIPS state code and the last five are the code for the city or place within the state or possession.
Congressional District	Two character state abbreviations used by the United States Postal Service followed by a dash then number (e.g. IA-1, IA-2, IA-3, IA-4, etc.)
Watershed HU08	A national standard hierarchical system for watersheds based on surface hydrologic features developed by the U.S. Geological Survey. Eight digit hydrologic unit code represents a subbasin or cataloguing unit for watersheds.
Watershed HU10	A national standard hierarchical system for watersheds based on surface hydrologic features developed by the U.S. Geological Survey. Ten digit hydrologic unit code represents a watershed.
Watershed HU12	A national standard hierarchical system for watersheds based on surface hydrologic features developed by the U.S. Geological Survey. Twelve digit hydrologic unit code representing a subwatershed.
Census Tracts	Number given to geographic region defined for the purpose of taking a decennial census. Usually coincides with the limits of cities, towns or other administrative areas and several tracts commonly exist within a county.
Census Block Groups	Number given to subdivision of a census tract.
Census Blocks	Number given to subdivision of census block group.
Latitude	North/south coordinate of a point in decimal degrees with + for North, and – for South. Values must be between -180 and 180.
Longitude	East/west coordinate of a point in decimal degrees with + for East, and – for West. Values must be between -180 and 180.
School District Numbers	Four-digit code assigned by the Iowa Department of Education for public school districts (or local education agencies) in Iowa.

Data Type	Definition
School AEA Numbers	Two-digit code assigned by the Iowa Department of Education for Area Education Agencies in Iowa.
Political Township Number	Unique identifier available in the Department of Natural Resource's shapefile for political townships.
Iowa Levy Authority Code	Code given by the Iowa Department of Management to entities authorized to collect property taxes.
State Department Number	Three-digit numeric code assigned to state departments within the accounting and budget systems.
State Business Unit	Four digit/character alpha numeric code assigned to sub-units of state agencies with the state accounting and budget systems.
State Fund Code	Four digit/character alpha numeric code assigned to the state general fund and over 700 other funds within the state accounting and budget systems.
State Appropriation Code	Three digit/character alpha numeric code (except for Fund Only, which is "0000") assigned to spending authorized by the Legislature, either from the State General Fund or from the other funds established by constitution or by legislative authorization.
State Object Class	Three-digit numeric code which categorizes expenses.
State Revenue Class	Three-digit numeric code identifying a revenue class, which categorizes revenue sources.
Web URL/Link	Provides a link starting with http or https.
Data.iowa.gov Node	The number assigned to content within data.iowa.gov.
BASE64 Encoded Image	Encoded images in base64 string format for usage in CSS, XHTML, XML and more.